

The Importance of Kernels for Performance Portability,

Or, How I Learned to Stop Looping and Love the Kernel

Tom Scogland, David Richards

September, 2020



The Volta Compute Unit

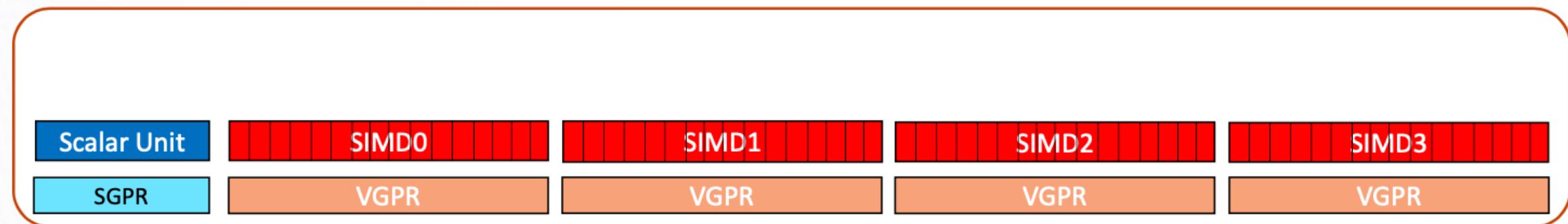
Major Features include:

- ▶ New mixed-precision Tensor Cores purpose-built for deep learning matrix arithmetic, delivering 12x TFLOPS for training, compared to GP100, in the same power envelope
- ▶ 50% higher energy efficiency on general compute workloads
- ▶ Enhanced high performance L1 data cache
- ▶ A new SIMT thread model that removes limitations present in previous SIMT and SIMD processor designs



NVIDIA TESLA V100 GPU ARCHITECTURE – whitepaper WP-08608-001_v1.1

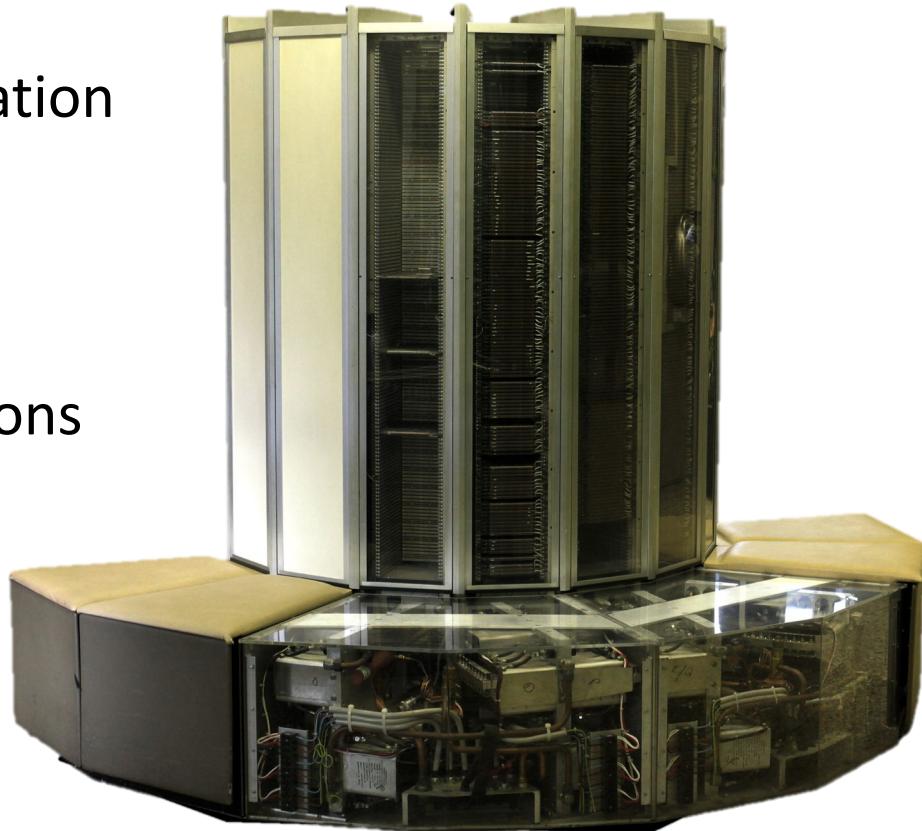
The GCN Compute Unit (CU)



- SIMD Units
 - 4x SIMD vector units (each 16 lanes wide)
 - 4x 64KB (256KB total) Vector General-Purpose Register (VGPR) file
 - Maximum of 256 registers per SIMD – each register is 64x 4-byte entries
 - Instruction buffer for 10 wavefronts on each SIMD unit
 - Each wavefront is local to a single SIMD unit, not spread among the four (more on this in a moment)

This is all Seymour's fault*

- Vector machines begat the SIMD machines of today, even if they were pipelined
- Early compilers either added explicit extensions for vectorization
 - Vector LRLTRAN
 - Vectra
 - Etc.
- Or vectorized sequential loops and ultimately array expressions descended from the extensions
- Trouble was, vectorizing sequential code is *hard*



*It's not really Seymour Cray's fault. He didn't invent vector processing. But he did create the successful vector processor.

The “Killer Micros” soon gave us SIMD vectors

- SIMD instruction sets were widely introduced in mid-90's (MMX, 3DNow, SSE, etc.)
- Instead of pipelining, early microprocessor SIMD:
 - Added wide SIMD registers
 - Processed elements in those registers in parallel (usually)
 - Had little support for cross-lane operations (reduction, swizzle, gather, scatter)
- Heavily reliant on data contiguous in memory
- Sequential code ~~still hard~~ even harder to vectorize
 - Relied heavily on programmers to manually vectorize
 - Compilers are notoriously conservative
 - Slow code is better than incorrect code
 - C is more popular than fortran by this point, C is much harder to auto-vectorize



Many programmers ignored SIMD units for many years because they were just so hard to use

SIMD code can be a nightmare to write and maintain

Original Code

```
void update_jGate(double dt, int nCells, double *VM,
double *g, double *mhu_a, double *tauR_a)
{
    int gateIndex=2;
    int mhu_l= 7 ;
    int mhu_m= 9;
    int taur_l= 1;
    int taur_m=13;
    double tauRdt_a[taur_m];
    for (int j=taur_m-1;j>=0;j--)    tauRdt_a[j] = tauR_a[j]*dt;
    int mhu_k = mhu_m+mhu_l-1;
    int taur_k = taur_m+taur_l-1;
    for (int ii=0;ii<nCells;ii++)
    {
        double x = VM[ii];
        double sum1=0;
        double sum2=0;
        double sum3=0;
        for (int j=mhu_m-1;j>=0 ;j--)sum1 = mhu_a[j] + x*sum1;
        for (int j=mhu_k ;j>=mhu_m;j--)sum2 = mhu_a[j] + x*sum2;
        for (int j=taur_m-1;j>=0 ;j--)sum3 = tauRdt_a[j] + x*sum3;
        double tauRdt= sum3;
        double mhu= sum1/sum2;
        g[ii] += mhu*tauRdt - g[ii]*tauRdt;
    }
}
```

27 lines of code

After SIMDization

```
void update_jGate(double dt, int nCells, double *VM,
double *g, double *mhu_a, double *tauR_a)
{
    vector4double v_xa, sum1,sum2,sum3
    .
    .
    .
    v_sum1a = vec_madd(v_xa, v_sum1a, v_mhu_A2);
    v_sum1a = vec_madd(v_xa, v_sum1a, v_mhu_A1);
    v_sum2a = vec_madd(v_xa, v_sum2a, v_mhu_B2);
    v_sum2a = vec_madd(v_xa, v_sum2a, v_mhu_B1);
    v_sum3a = vec_madd(v_xa, v_sum3a, v_tauRdt_C2);
    v_sum3a = vec_madd(v_xa, v_sum3a, v_tauRdt_C1);
    sum1[ii] = mhu_a[0]      + VM[ii]*sum1[ii];//BODYCOM0
    sum2[ii] = mhu_a[10]     + VM[ii]*sum2[ii];//BODYCOM0
    sum2[ii] = mhu_a[9]      + VM[ii]*sum2[ii];//BODYCOM0
    sum3[ii] = tauRdt_a[13] + VM[ii]*sum3[ii];//BODYCOM0
    sum3[ii] = tauRdt_a[12] + VM[ii]*sum3[ii];//BODYCOM0
    v_mhu_A1 = vec_lds(0, &mhu_a[0]);//BODYCOM0
    v_mhu_B1 = vec_lds(0, &mhu_a[9]);//BODYCOM0
    v_mhu_B2 = vec_lds(0, &mhu_a[10]);//BODYCOM0
    v_tauRdt_C1 = vec_lds(0, &tauRdt_a[12]);//BODYCOM0
    v_tauRdt_C2 = vec_lds(0, &tauRdt_a[13]);//BODYCOM0
    .
    .
    .
}
```

537 lines of code

Even Simple Code, Like DAXPY

Sequential

```
for (int i = 0; i < n; i++)
    y[i] = a * x[i] + y[i];
```

After SIMDization

```
while (n >= 2) {
    n -= 2;
    _mm_store_pd(y + n, _mm_add_pd(_mm_load_pd(y + n),
        _mm_mul_pd(_mm_load_pd(x + n), _mm_load1_pd(&a))));
```

Even the simplest code becomes tricky, and non-portable, with simd intrinsics

GPUs gave us wider vectors but also a new programming model

- Step 1 of GPU porting is converting loops to kernels

Sequential

```
void daxpy(int n, double a, double *x, double *y){  
    for (int i = 0; i < n; i++)  
        y[i] = a * x[i] + y[i];  
}
```

CUDA

```
__global__ void daxpy_knl(int n, double a,  
                           double *x, double *y){  
    int i = threadIdx.x + blockIdx.x * blockDim.x;  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

OpenCL

```
__kernel void daxpy(int n, double a,  
                     __global double *src, __global double *dst)  
{  
    int i = get_global_id(0);  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

Brook+

```
kernel void daxpy(int n, double a,  
                  double x<>, out double y<>) {  
    int i = instance().x;  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

GPUs gave us wider vectors but also a new programming model

- Step 1 of GPU porting is converting loops to kernels

Sequential

```
void daxpy(int n, double a, double *x, double *y){  
    for (int i = 0; i < n; i++)  
        y[i] = a * x[i] + y[i];  
}
```

CUDA

```
__global__ void daxpy_knl(int n, double a,  
                           double *x, double *y){  
    int i = threadIdx.x + blockIdx.x * blockDim.x;  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

OpenCL

```
__kernel void daxpy(int n, double a,  
                    __global double *src, __global double *dst)  
{  
    int i = get_global_id(0);  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

Brook+

```
kernel void daxpy(int n, double a,  
                  double x<>, out double y<>) {  
    int i = instance().x;  
    if (i < n) y[i] = a * x[i] + y[i];  
}
```

It is easy to overlook the significant differences between kernels and loops
(and the execution model that goes with them).

GPU Kernels *Always* Vectorize

- A GPU model has no scalar context
- A CPU model has a native scalar context but allows vector instructions in that context
- Kernel code **always** vectorizes
 - For some meaning of vectorize
 - No compiler intelligence necessary

```
__global__ void daxpy_knl(int n, double a,
                           double *x, double *y)
{
    int i = threadIdx.x + blockIdx.x * blockDim.x;
    if (i < n) y[i] = a * x[i] + y[i];
}
```

What instructions does gcc generate for a basic loop?

```
void naive_memcpy(char * dst, char * src, size_t n){  
    for (int i=0; i<n; i++)  
        dst[i] = src[i];  
}
```

```
1  exmemcpy(char*, char*, unsigned long):  
2      mov     rcx, rdx  
3      test    rdx, rdx  
4      je     .L8  
5      xor     eax, eax  
6  .L3:  
7      movzx   edx, BYTE PTR [rsi+rax]  
8      mov     BYTE PTR [rdi+rax], dl  
9      inc     rax  
10     cmp    rax, rcx  
11     jne     .L3  
12  .L8:  
13      ret
```

*With reasonable flags: -O2 –march=haswell -fopenmp

One operation, one byte at a time (move, test, branch)

What Happens With an OMP parallel for Loop?

```
void myMemcpy4(char * dst,
               char * src,
               size_t n) {
    #pragma omp parallel for
    for(int i = 0; i < n; ++i)
        dst[i] = src[i];
}
```

Loop over chunks

Déjà vu. Same
serial code
executed for a
chunk



```
call    omp_get_thread_num
mov     ecx, eax
mov     rax, QWORD PTR [rbx+16]
cdq
idiv   ebp
cmp     ecx, edx
jl     .L2

.L5:
imul   ecx, eax
add    edx, ecx
add    eax, edx
cmp    edx, eax
jge    .L7
mov    rdi, QWORD PTR [rbx+8]
mov    rsi, QWORD PTR [rbx]
movsx  rdx, edx

.L4:
movzx  ecx, BYTE PTR [rdi+rdx]
mov    BYTE PTR [rsi+rdx], cl
inc    rdx
cmp    eax, edx
jg     .L4

.L7:
add    rsp, 8
pop    rbx
pop    rbp
ret

.L2:
inc    eax
xor    edx, edx
jmp    .L5
```

What PTX does nvcc generate for a similar kernel?

```
global__ void mymemcpy(char *dest, char *src, size_t n) {  
    int tid = blockIdx.x * blockDim.x + threadIdx.x;  
    if (tid < n)  
        dest[tid] = src[tid];  
}
```

- Same memcpy as a cuda kernel
- PTX looks similar to sequential loop
- **But these are *vector* instructions across warps**
- There are no truly sequential instructions here

```
8      ld.param.u64    %rd1, [_Z8memcpyPcS_m_param_0];  
9      ld.param.u64    %rd2, [_Z8memcpyPcS_m_param_1];  
10     ld.param.u64    %rd3, [_Z8memcpyPcS_m_param_2];  
11     mov.u32          %r2, %ctaid.x;  
12     mov.u32          %r3, %ntid.x;  
13     mov.u32          %r4, %tid.x;  
14     mad.lo.s32       %r1, %r3, %r2, %r4;  
15     cvt.s64.s32     %rd4, %r1;  
16     setp.ge.u64      %pl, %rd4, %rd3;  
17     @%pl bra         BB0_2;  
18  
19     cvta.to.global.u64  %rd5, %rd2;  
20     add.s64           %rd7, %rd5, %rd4;  
21     ld.global.u8      %rs1, [%rd7];  
22     cvta.to.global.u64  %rd8, %rd1;  
23     add.s64           %rd9, %rd8, %rd4;  
24     st.global.u8      [%rd9], %rs1;  
25  
26     ret;
```

Making a kernel behave sequentially

- If we *want* sequential behavior it has to be **explicit**
- Parallel and *unsequenced* by default

```
_global_ void sequential_mc(char *dst, char *src, size_t n){  
    if (thread_id == 0) {  
        for(int i = 0; i < n; ++i)  
            dst[i] = src[i]  
    }  
}
```

- Far more obvious this is a Bad Idea™
- Generates **74** PTX instructions to the 17 in the unsequenced

GPU models aren't the only way to write kernels: OpenMP

```
void mc_omp_simd(char * dest, char * src, size_t n) {  
    // Copy contents of src[] to dest[]  
#pragma omp simd  
    for(int i = 0; i < n; ++i)  
        dest[i] = src[i];  
}
```

- Now unsequenced (thanks to **simd** directive)
- Generates **three** loops
 - 256-bit vectors
 - 128-bit vectors
 - Individual bytes
- Ties them together to do the most efficient thing for a length

```
myMemcpy3(char*, char*, unsigned long):  
    mov    r8d, edx  
    test   edx, edx  
    jle    .L24  
    lea    r9d, [rdx-1]  
    cmp    r9d, 30  
    jbe    .L20  
    mov    ecx, edx  
    xor    eax, eax  
    shr    ecx, 5  
    sal    rcx, 5  
.L15:  
    vmovdqu ymm0, YMMWORD PTR [rsi+rax]  
    vmovdqu YMMWORD PTR [rdi+rax], ymm0  
    add    rax, 32  
    cmp    rax, rcx  
    jne    .L15  
    mov    ecx, edx  
    and   ecx, -32  
    mov    eax, ecx  
    cmp    edx, ecx  
    je     .L25  
    vzeroupper  
.L14:  
    sub    r9d, ecx  
    sub    edx, ecx  
    cmp    r9d, 14  
    jbe    .L17  
    vmovdqu xmml, XMMWORD PTR [rsi+rax]  
    vmovdqu XMMWORD PTR [rdi+rax], xmml  
    mov    ecx, edx  
    and   ecx, -16  
    add    eax, ecx  
    cmp    edx, ecx  
    je     .L24  
.L17:  
    cdqe  
.L19:  
    movzx  edx, BYTE PTR [rsi+rax]  
    mov    BYTE PTR [rdi+rax], dl  
    inc    rax  
    cmp    r8d, eax  
    jg    .L19  
.L24:  
    ret  
.L20:  
    xor    ecx, ecx  
    xor    eax, eax  
    jmp    .L14  
.L25:  
    vzeroupper  
    ret
```

GPU models aren't the only way to write kernels: C++17/RAJA/Kokkos

```
void myMemcpy6(char * dest, char * src, size_t n) {
    auto r = RAJA::RangeSegment(0, n);
    std::for_each(
        std::execution::par_unseq, r.begin(), r.end(),
        [=](size_t i) {
            dest[i] = src[i];
        });
}
```

- Similarly vectorized
- Only two loops this time, 256-bit and per-byte

```
std::enable_if<_pstl::execution::v1::is_execution_policy<std::remove_cv<std::remove_reference<
```

```
        mov    r8, rdi
        cmp    rsi, rdi
        ja    .L37
        mov    rax, rdi
        mov    rdi, rsi
        sub    rax, rsi
        sub    rdi, r8
.L38:
        test   rax, rax
        jns    .L50
        test   rdi, rdi
        mov    r9d, 1
        cmovg r9, rdi
        cmp    rdi, 31
        jle    .L46
        mov    rsi, r9
        add    rcx, r8
        add    rdx, r8
.L41:
        vmovdqu ymm0, YMMWORD PTR [rcx+rax]
        vmovdqu YMMWORD PTR [rdx+rax], ymm0
        add    rax, 32
        cmp    rax, rsi
        jne    .L41
        mov    rax, r9
        and    rax, -32
        and    r9d, 31
        je     .L51
        vzeroupper
.L45:
        movzx  esi, BYTE PTR [rcx+rax]
        mov    BYTE PTR [rdx+rax], sil
        inc    rax
        cmp    rdi, rax
        jg    .L45
.L50:
        ret
.L37:
        mov    rdi, rsi
        mov    rax, r8
        sub    rdi, r8
        sub    rax, rsi
        jmp    .L38
.L46:
        xor    eax, eax
        add    rcx, r8
        add    rdx, r8
        jmp    .L45
.L51:
```

C++17 execution policies can express parallel unsequenced kernels as well

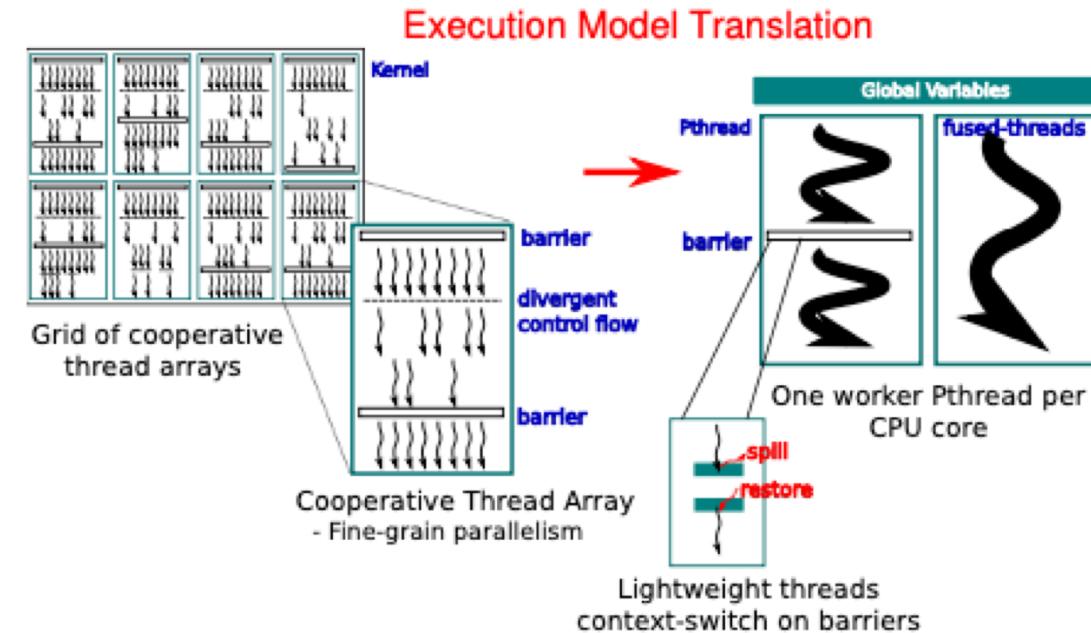
C++17 and RAJA give flexibility through *policies*

- Serial sequenced or unsequenced:
 - RAJA::simd or std::unseq: unsequenced kernel, vector safe
 - RAJA::loop or std::seq: sequenced, order matters, compiler must prove or provide safety
- Parallelism is another axis entirely:
 - RAJA::omp_parallel_for_exec or std::par: run in parallel but still *sequenced in each thread*
 - RAJA::omp_parallel_for_simd_exec or std::par_unseq: both parallel and unsequenced execution are safe

Logically a “kernel”, policy determines strength of guarantees on the spectrum from loop to unsequenced parallel kernel

Kernels that use team/block synchronization have weaknesses too

- Full context (stack, registers) must persist for each logical thread across the synchronization
- Huge context switch and flushing/loading cost to make progress for many logical threads (500+)
- This is why Apple OpenCL for CPU only supported **one thread** per workgroup
- If there are no barriers (syncthreads, workgroup_sync, etc.) there's no cost, if there is CPU may suffer while GPU does not



Gregory Diamos, Andrew Kerr, Sudhakar Yalamanchili, and Nathan Clark. "Ocelot: A Dynamic Compiler for Bulk-Synchronous Applications in Heterogeneous Systems." The Nineteenth International Conference on Parallel Architectures and Compilation Techniques. September 2010.

Conclusions

- Eliminating sequential scalar loops from your code and writing in “kernel” or an unsequenced parallel form is inherently Performant, Portable, and Productive across vendors and architectures
- **But it's not a silver bullet**
 - Data organization and motion are still a concern
 - User must prove the safety and lack of races or data ordering issues, compiler is absolved of dependencies between iterations, lanes, or threads
 - Synchronization can have widely varying costs at different scopes on different architectures
- Overall, learn to love the kernel





**Lawrence Livermore
National Laboratory**

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

We need to keep three promises in this talk

- 1. Explain why the kernel programming model produces code that always vectorizes
 - Slides 1-6 give background and explain the vectorization challenge
- 2. Compare RAJA `forall`, Kokkos `parallel_for`, and OpenMP `omp loop` to `omp for`
 - `Omp for` is a set of sequential loops in parallel
 - This is where your code examples come in
- 3. Discuss weaknesses of kernels
 - And why weaknesses are outweighed by advantages
- Conclusion: Eliminating sequential scalar loops from your code and writing in kernel form is inherently Performant, Portable, and Productive across vendors and architectures
 - But it's not a silver bullet. You still need to worry about data organization and motion

Cutting room floor

- We could talk about array programming
 - Fortran 90 looks like an example, but really isn't.
 - We would talk to John Levesque about why Fortran vector notation isn't nearly as nice as you think it might be for OpenMP programming.
- 925-577-9005

Godbolt links for examples



**Lawrence Livermore
National Laboratory**